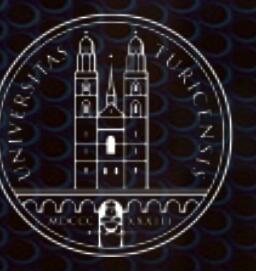


BIO392

Bioinformatics of Genome Variations

Genomes: Core of "Personalized Health" & "Precision Medicine"

Michael Baudis **UZH SIB**
Computational Oncogenomics



University of
Zurich^{UZH}

Variant - Disease Knowledge Bases

VARIANT RESOURCES FOR CANCER GENOMICS

Resource name	Primary institute	Constituent Knowledge base	Cancer focused	Therapeutic evidence	Predisp. evidence	Diagnostic evidence	Prognostic evidence	Variant emphasis	URL
Cancer Genome Interpreter (CGI)	Institute for Research in Biomedicine, Barcelona, Spain	x	x	x				Somatic	https://www.cancergenomeinterpreter.org/home
Clinical Interpretation of Variants in Cancer (CIViC)	Washington University School of Medicine (WashU)	x	x	x	x	x	x	All variants	www.civicdb.org
JAX Clinical Knowledgebase (CKB)	The Jackson Laboratory	x	x	x	x	x	x	Somatic	https://ckb.jax.org/
Molecular Match	Molecular Match	x	x	x			x	Somatic	https://app.molecularmatch.com/
OncoKB	Memorial Sloan Kettering Cancer Center	x	x	x				Somatic	http://oncokb.org/#/
Precision Medicine Knowledgebase (PMKB)	Weill Cornell Medical College	x	x	x	x	x	x	Somatic	https://pmkb.weill.cornell.edu/
BRCA exchange	GA4GH	x	x		x			Germline	http://brcaexchange.org/
Cancer Driver Log (CanDL)	Ohio State University (OSU) / James Cancer Hospital		x	x				Somatic	https://cndl.osu.edu/
Gene Drug Knowledge Database	Synapse		x	x		x	x	Somatic	https://www.synapse.org/#!Synapse:syn2370773/wiki/62707
MatchMiner	Dana-Farber Cancer Institute		x					Somatic	http://bcb.dfci.harvard.edu/knowledge-systems/
COSMIC Drug Resistance Curation	Wellcome Trust Sanger Institute		x	x				Somatic	http://cancer.sanger.ac.uk/cosmic/drug_resistance
My Cancer Genome	Vanderbilt University		x	x		x	x	Somatic	https://www.mycancergenome.org/
NCI Clinical Trials	National Cancer Institute of the National Institutes of Health		x					Somatic	www.cancer.gov/about-cancer/treatment/clinical-trials
Personalized Cancer Therapy Database	MD Anderson Cancer Center		x	x	x	x	x	Somatic	https://pct.mdanderson.org/#/home
ClinGen Knowledge Base	ClinGen				x			Germline	https://www.clinicalgenome.org/resources-tools/
ClinVar	National Center for Biotechnology Information (NCBI)			x	x			All variants	http://www.ncbi.nlm.nih.gov/clinvar/
Pharmacogenomics Knowledgebase (PharmGKB)	Stanford University			x				Germline	https://www.pharmgkb.org/
The Human Gene Mutation Database (HGMD)	Institute of Medical Genetics in Cardiff				x			Germline	http://www.hgmd.cf.ac.uk

CANCER VARIANT KNOWLEDGE BASES

- ▶ cancer variant knowledge databases report evidences for disease association (causative, therapeutic targets...)
 - ▶ data selection is driven by arbitrary observations and sample selections
 - ▶ limited overlap of reported variant associations is evidence for large gaps in knowledge

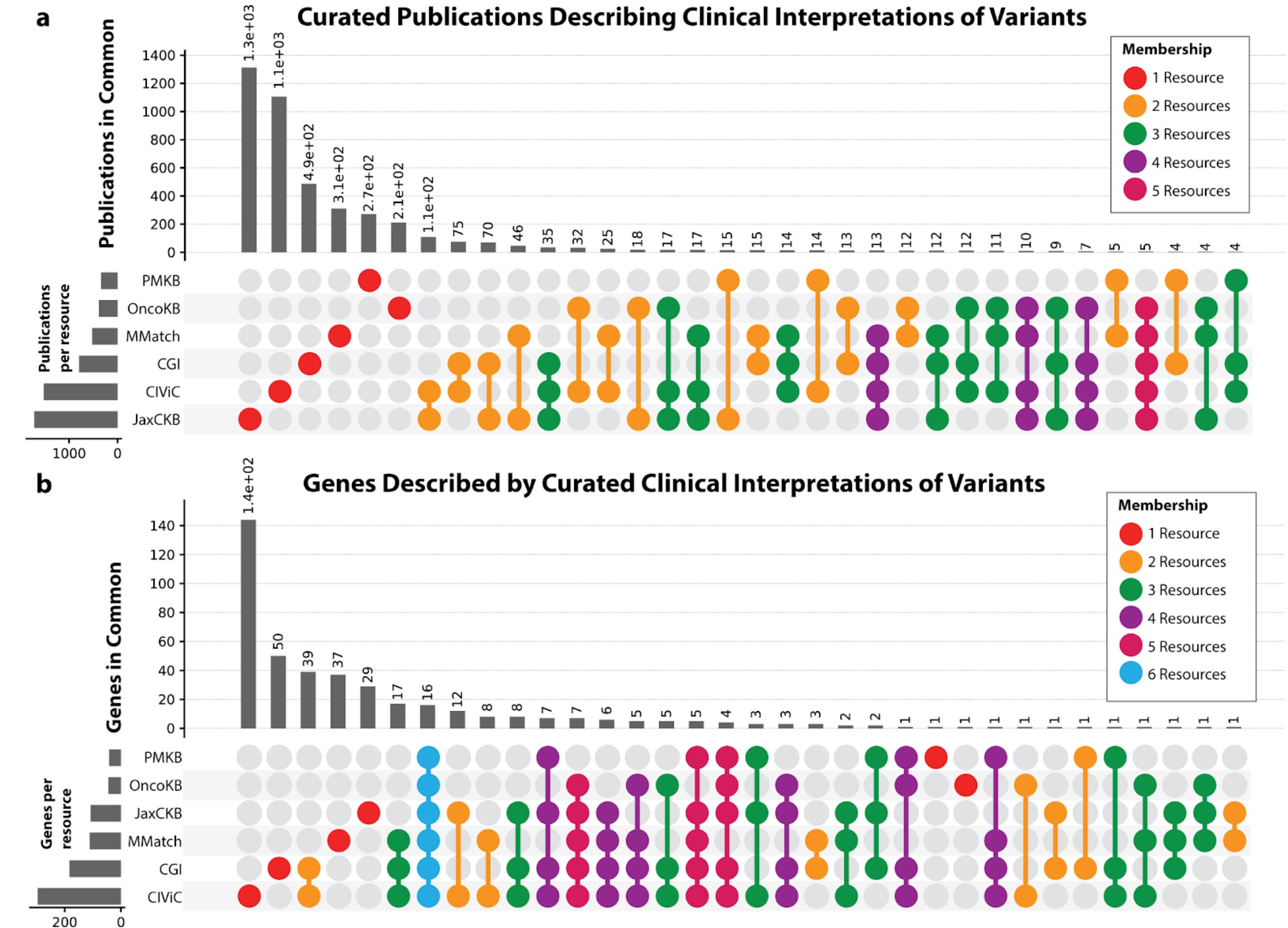


Figure S2 - Knowledgebase overlap

(a) Upset plot of publications supporting clinical interpretations of variants. the overwhelming majority of publications are observed in only 1 of 6 resources. **(b)** Upset plot of genes described by clinical interpretations of variants. Compared to other interpretation elements, genes are much more commonly shared between resources.

CANCER VARIANT KNOWLEDGE BASES

- ▶ cancer variant knowledge databases report evidences for disease association (causative, therapeutic targets...)
- ▶ data selection is driven by arbitrary observations and sample selections
- ▶ limited overlap of reported variant associations is evidence for large gaps in knowledge

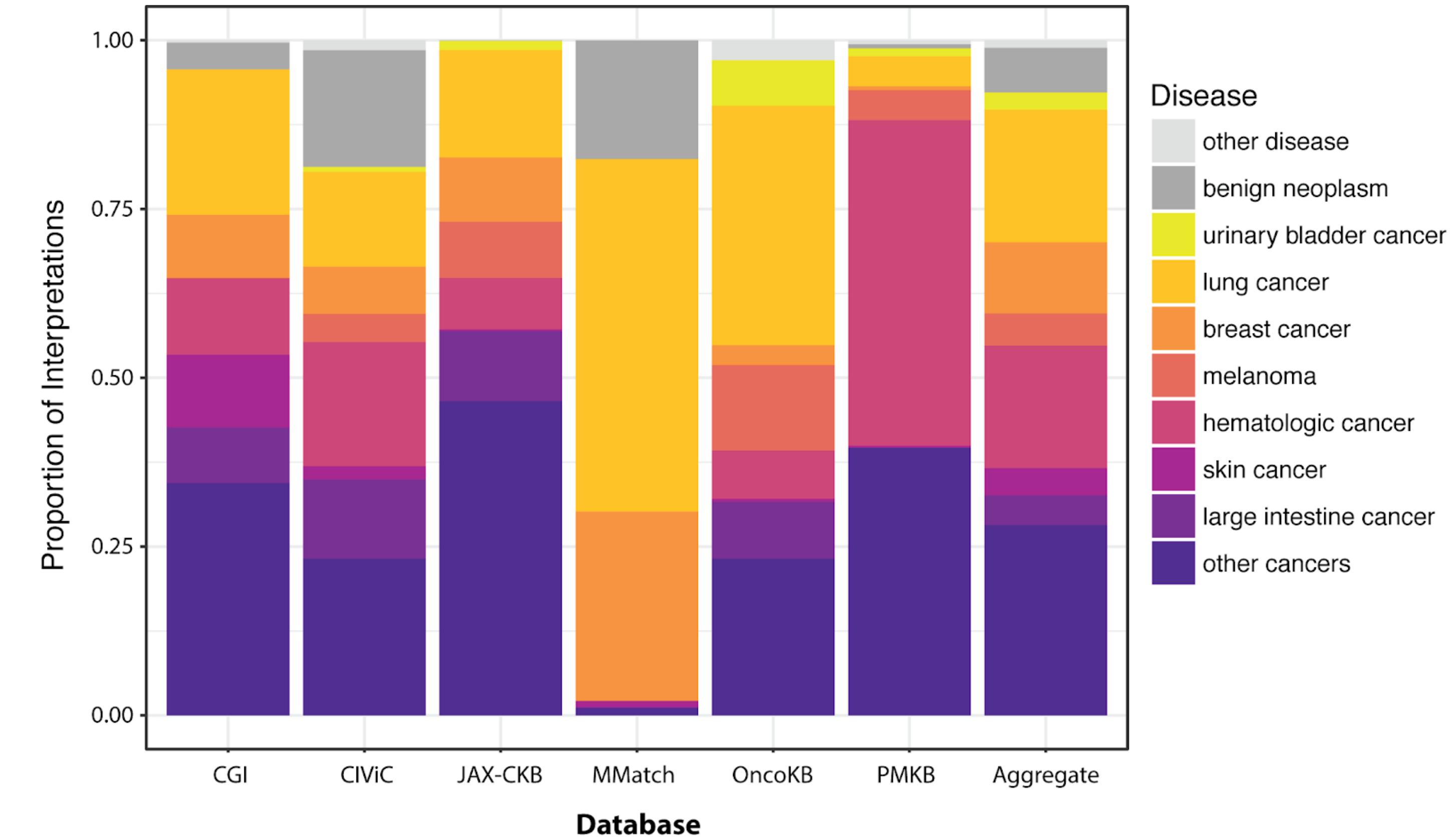


Figure S3 - Knowledgebase disease enrichment

Relative distribution of interpretations describing diseases across the VICC resources. Several resources are strongly enriched for one or more diseases compared to the entire dataset (see related **Table S8**).

RESOURCES FOR GENOMICS: CLINGEN

- ▶ "The Genomic Variant WG brings together representatives from the Sequence and Structural Variant communities for focused discussions on resolving discrepancies in variant interpretation and creating consistent curation guidelines."
- ▶ Interpreted genome variants with disease association

The screenshot shows the ClinGen Clinical Genome Resource website. At the top right is a search bar with the placeholder "Search our Knowledge Base for genes and diseases..." and a magnifying glass icon. Below the search bar are navigation links: About ClinGen, Working Groups, Resources, GenomeConnect, Share Your Data (highlighted in blue), and Curation Activities. The main banner features a blue background with a blurred image of laboratory glassware and a DNA sequence. The text "Defining the clinical relevance of genes & variants for precision medicine and research..." is displayed. Below the banner are three large numbers: 1496 ClinGen Curated Genes, 31 Expert Groups, and 10446 Expert Reviewed Variants in ClinVar, each with a corresponding icon. To the right is a "Knowledge Base Search" button with a magnifying glass icon. Below the banner, the tagline "Sharing Data. Building Knowledge. Improving Care." is followed by a description of ClinGen's mission. Six call-to-action boxes are arranged in a grid at the bottom:

- ClinGen-ClinVar Partnership (blue circular icon with DNA)
- How to share genomic & health data (blue circular icon with DNA and arrows)
- Learn about ClinGen curation activities (monitor icon with DNA)
- GenomeConnect Patient Registry (DNA helix icon)
- View ClinGen's Resources & Tools (laptop and smartphone icons)
- Get Involved (magnifying glass and notepad icon)

clinicalgenome.org

CANCER VARIANT KNOWLEDGE BASES: CIVIC

- ▶ "CIViC is a community-edited forum for discussion and interpretation of peer-reviewed publications pertaining to the clinical relevance of variants (or biomarker alterations) in cancer."

The screenshot shows the CIViC website interface. At the top, there is a navigation bar with links for About, Participate, Community, Help, FAQ, and Sign In/Sign Up. Below the navigation bar, the main content area is titled "GENE BRAF". On the left side of the main content, there is a detailed text block about BRAF mutations, sources (Li et al., 2009; Oncol. Rep. and Pakneshan et al., 2013, Pathology), and a "BRAF Variants & Variant Groups" section containing various variant categories and their icons. On the right side, there is a large blue-bordered box containing detailed information about the BRAF gene, including its name, Entrez symbol, aliases, chromosome location, protein domains, and pathways. A "View MyGene.info Details" button is located at the bottom right of this box. The overall design is clean and modern, with a dark header and light-colored content areas.

BRAF Variants & Variant Groups

AMPLIFICATION

599INST AGK-BRAF AKAP9-BRAF DELNVTAP F595L G464V G466A G466V G469A G469E G469R G469V G496A G596C G596R G596V

G606E intron 10 rearrangement intron 9 rearrangement K483M K601E KIAA1549-BRAF L505H L597Q L597R L597S L597V

MACF1-BRAF Fusion MUTATION N581S Non-V600 P731T PAPSS1-BRAF PPFIBP2-BRAF TRIM24-BRAF V600 V600_K601DELNSD

V600D V600E V600E AMPLIFICATION V600E+V600M V600E/K AMPLIFICATION V600K V600R WASFL-BRAF Fusion WILD TYPE

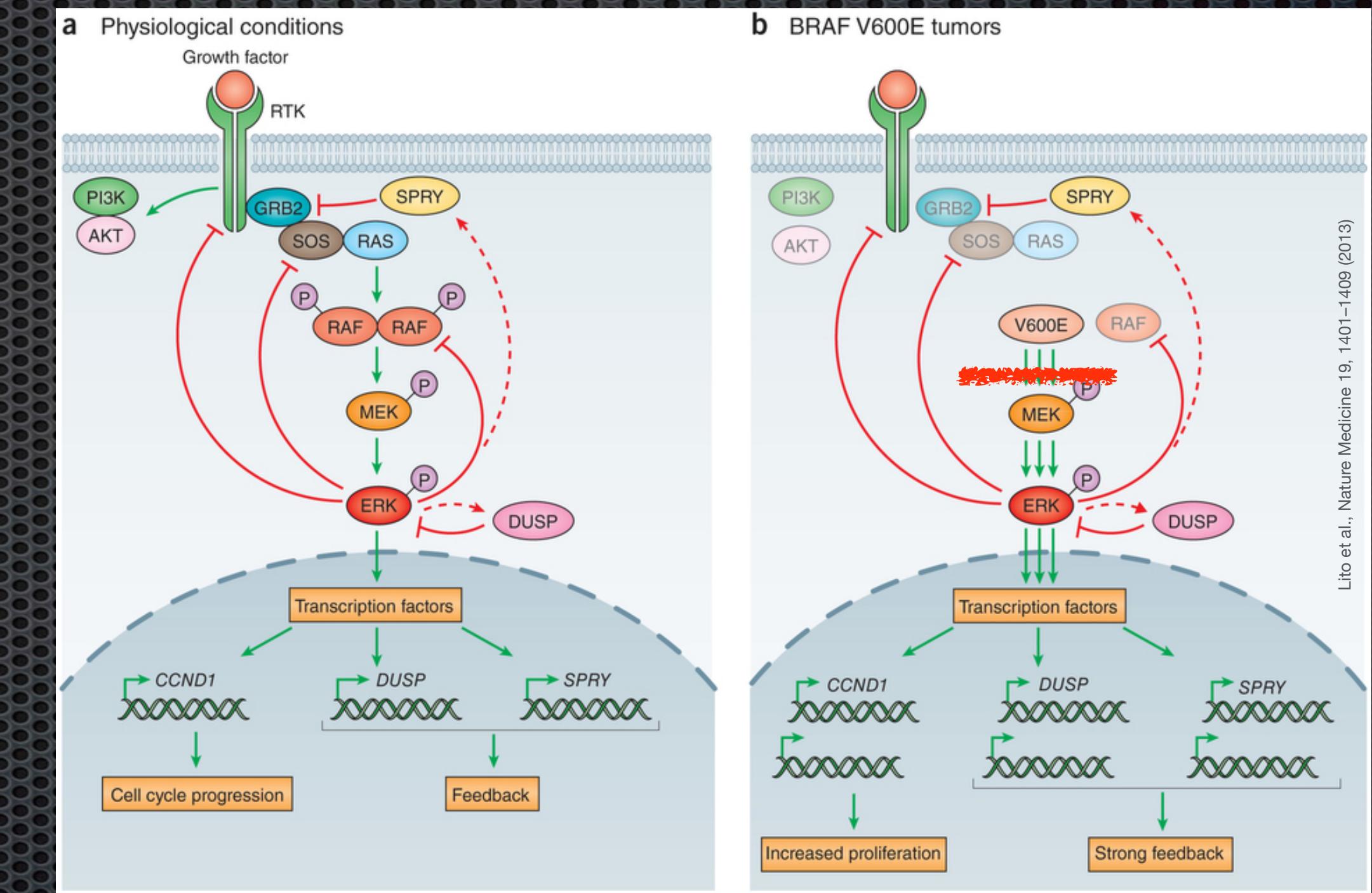
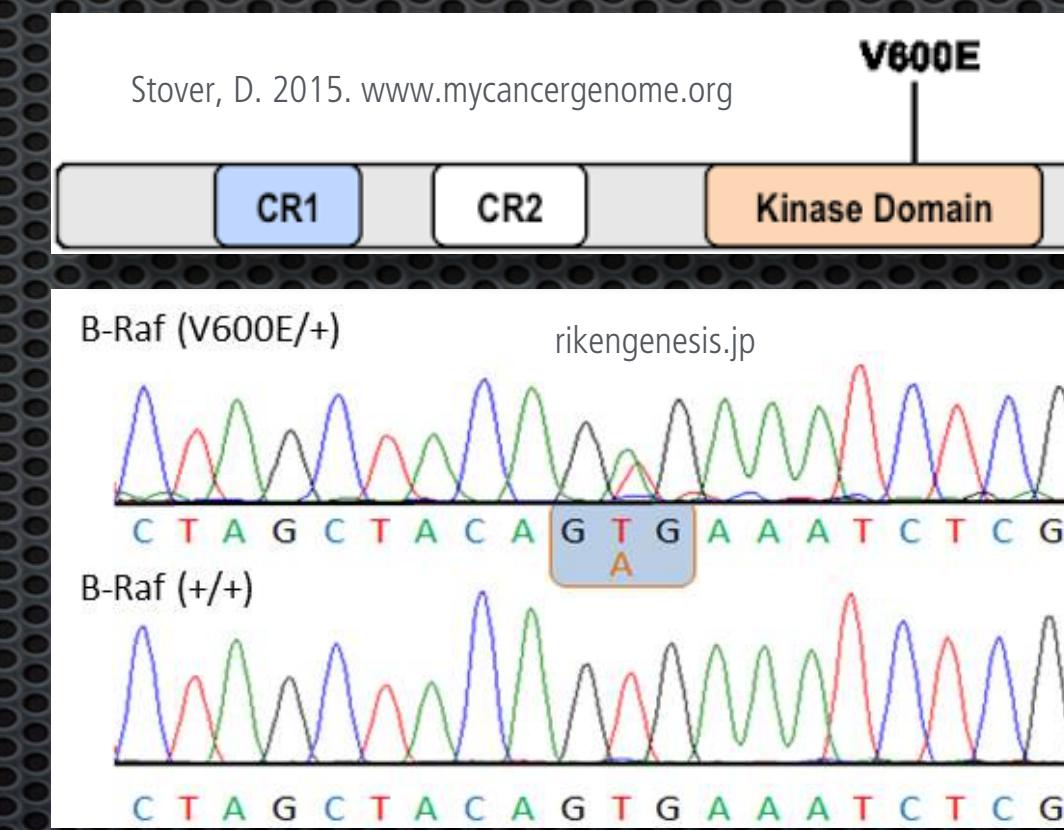
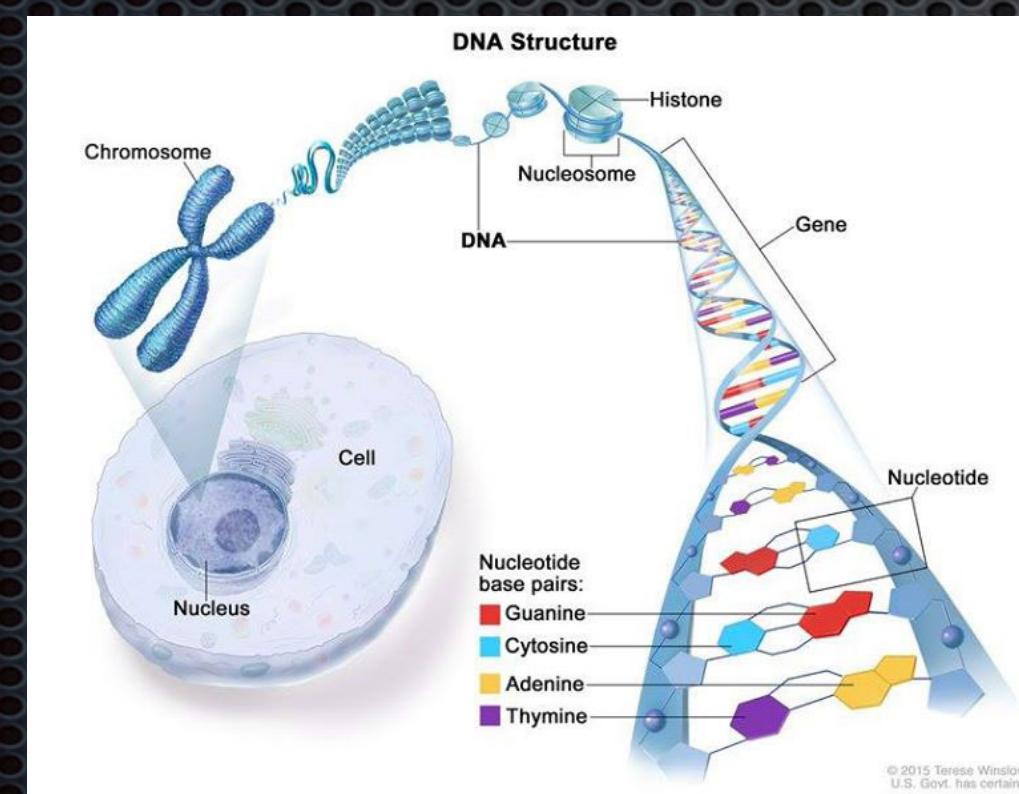
ZKSCAN1-BRAF

Name: B-Raf proto-oncogene, serine/threonine kinase
Entrez Symbol: BRAF Entrez ID: 673
Aliases: B-RAF1, B-raf, BRAF1, NS7, RAFB1
Chromosome: 7 Start: 140419127 End: 140624564 Strand: -1 (GRCh37)
Protein Domains: Diacylglycerol/phorbol-ester binding, Protein kinase C-like, phorbol ester/diacylglycerol-binding domain, Protein kinase domain, Protein kinase, ATP binding site, Protein kinase-like domain... (5 more)
Pathways: Intracellular Signalling Through Adenosine Receptor A2a and Adenosine, Intracellular Signalling Through Adenosine Receptor A2b and Adenosine, EGFR1, MAPK signaling pathway - Homo sapiens (human), ErbB signaling pathway - Homo sapiens (human)... (139 more)

View MyGene.info Details

BRAF V600E (c.1799T>A) Mutation Oncogene Activation by Single Nucleotide Alteration

- a single nucleotide exchange Thymidine > Adenine leads to continuous RAF based activation of the MEK-ERK pathway
- BRAF V600E is a frequent mutation in >50% of malignant melanomas, but also CRC, lung ADC ...
- pharmacologic block of B-Raf (e.g. through **Vemurafenib**)



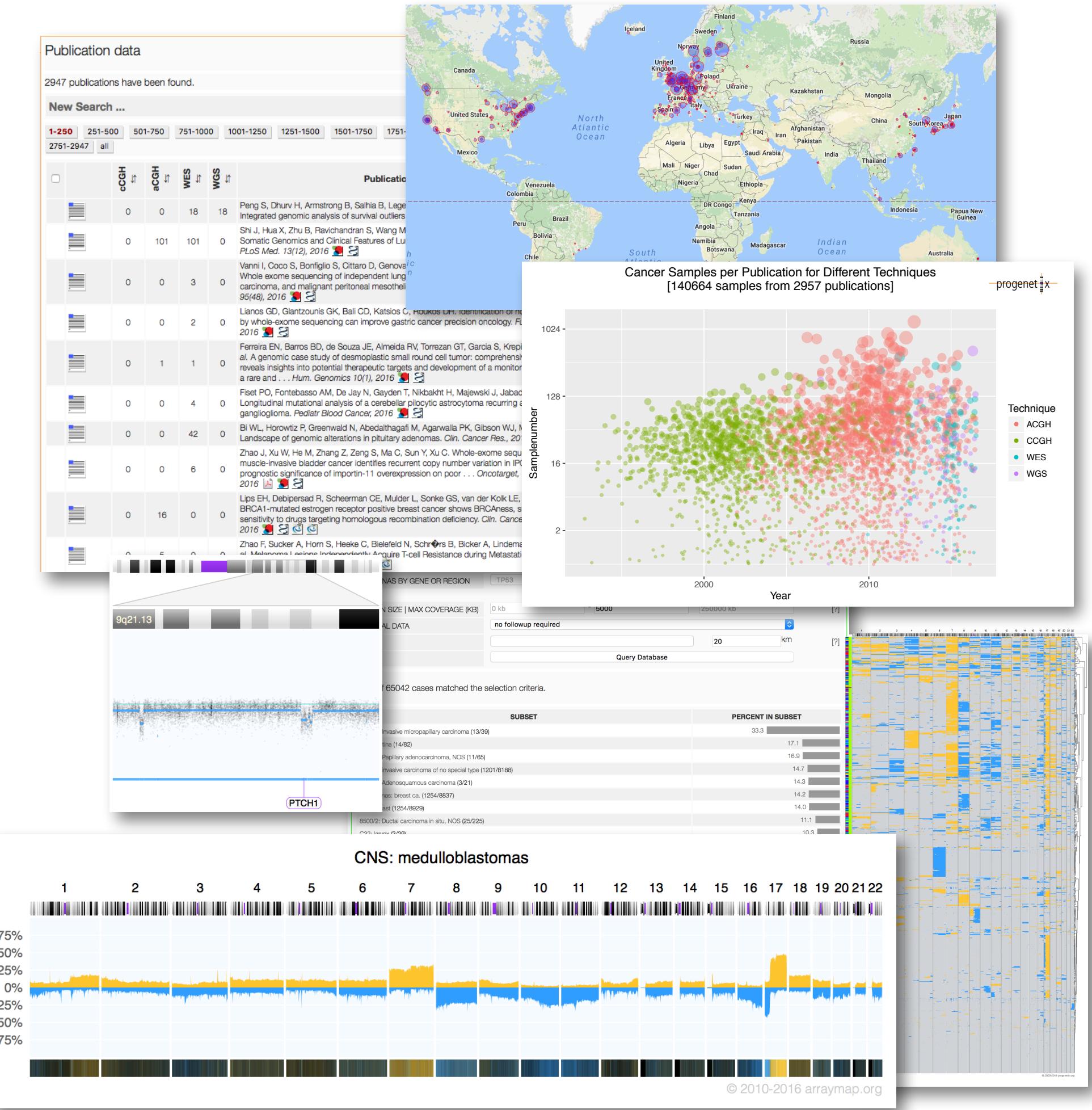
The BRAF V600E mutation leads to continuous phosphorylation of MEK, without the need for receptor based activation of the upstream pathway and loss of inhibitory feedback control.



Our contributions I: Cancer genome knowledge resources



- ▶ curated cancer genome publication resource (more than 3000 manually curated articles)
- ▶ article metadata and annotated genome profiles
- ▶ ontology mappings; clinical data where available
- ▶ epistemology: geographic and histologic sampling biases



- ▶ more than 60'000 array based genome profiles
- ▶ probe level, copy number, metadata
- ▶ completely open data access through web interface, downloads and API calls
- ▶ re-annotated metadata (diagnostic coding, basic clinical) for all samples

arrayMap

Resource for copy number variation data in cancer

arrayMap 

visualizing cancer genome array data @ arraymap.org

arrayMap is a curated reference database and bioinformatics resource targeting copy number profiling data in human cancer. The arrayMap database provides an entry point for meta-analysis and systems level data integration of high-resolution oncogenomic CNA data.

The current data reflects:

- 63060 genomic copy number arrays
- 763 experimental series
- 145 array platforms
- 141 ICD-O cancer entities
- 554 publications (Pubmed entries)

 **University of Zurich** ^{UZH}

Citation
User Guide
Registration & Licensing
People
External Links ↗

FOLLOW US ON [twitter](#)

 130.60.23.21

For the majority of the samples, probe level visualization as well as customized data representation facilitate gene level and genome wide data review. Results from multi-case selections can be connected to downstream data analysis and visualization tools, as we provide through our Progenetix project.

arrayMap is developed by the group "Theoretical Cytogenetics and Oncogenomics" at the Institute of Molecular Life Sciences of the University of Zurich.

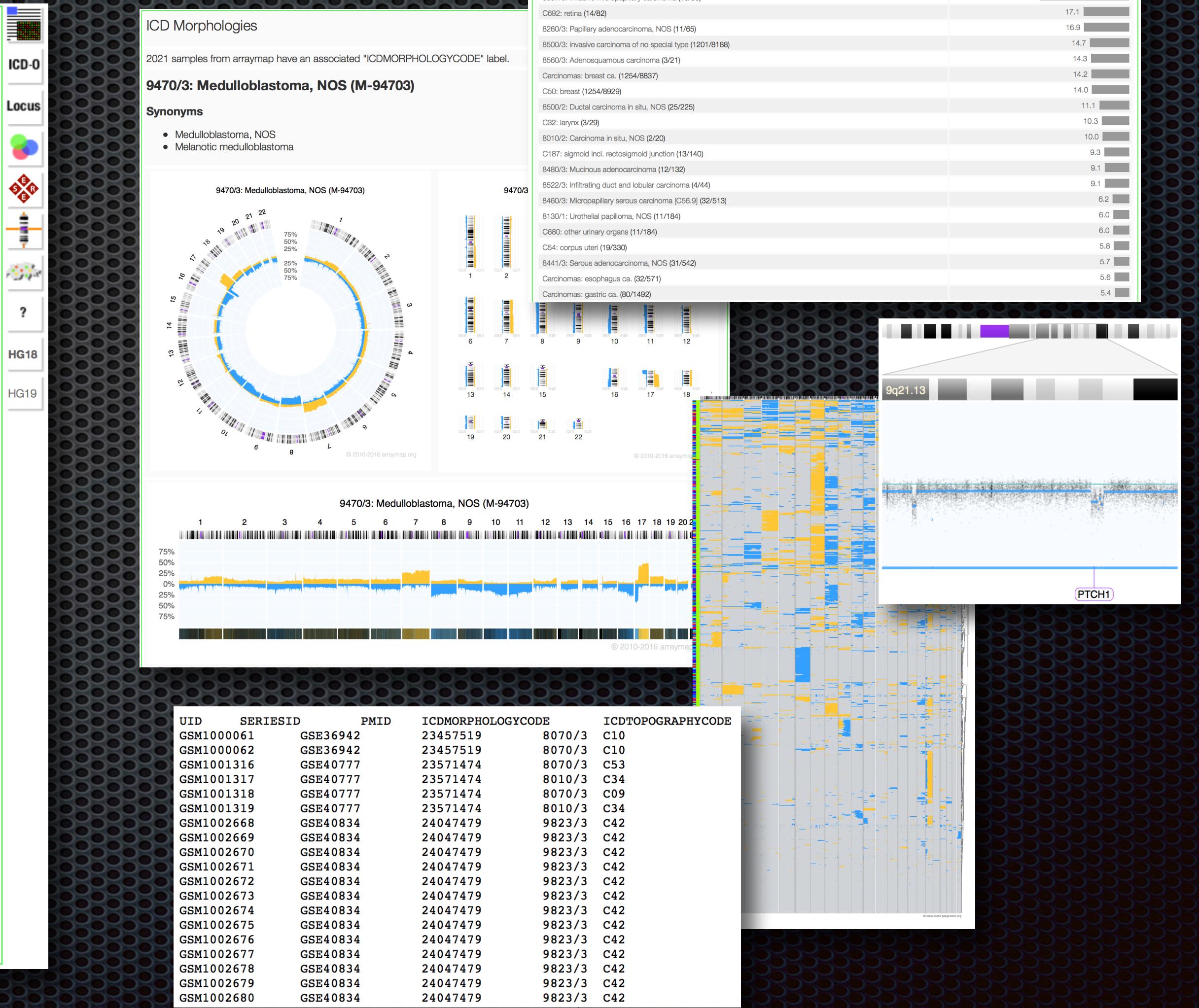
BRAIN TUMOURS	5653 samples ↗	[?]
BREAST CANCER	8329 samples ↗	[?]
COLORECTAL CANCER	3238 samples ↗	[?]
PROSTATE CANCER	991 samples ↗	[?]
STOMACH CANCER	1062 samples ↗	[?]

ARRAYMAP NEWS

- 2016-08-03: SVG graphics
- 2016-05-17: Transitioning to Europe PMC
- More news ...

Feel free to use the data and tools for academic research projects and other applications. If more support and/or custom analysis is needed, please contact Michael Baudis regarding a collaborative project or a special license.

© 2000 - 2016 Michael Baudis, refreshed Mon, 19 Sep 2016 10:20:09 GMT in 6.87s on server 130.60.240.68. No responsibility is taken for the correctness of the data presented nor the results achieved with the Progenetix tools.



arrayMap

Resource for copy number variation data in cancer

- Typical use: Gene CNA frequencies
- allows for (gene symbol supported) query for DUP/DEL overlapping a region
- size windowing ("focal" changes)
- no thorough statistics implemented yet (e.g. modeling overall CNA background)

arrayMap

Search Samples
Search Publications
Gene CNA Frequencies
User Data
Progenetix
 University of Zurich UZH
Citation
User Guide
Registration & Licensing
People

 130.226.87.158

Search for single or combined imbalances

FIND CNAS BY GENE OR REGION
Gene Symbol: TP53 [CDKN2A] 9:21957750-21984490:DEL [?]
CNA REGION SIZE (KB)
Minimal size: minimum CNA segment [?] Maximum size: 5000 [?]
SELECT CANCER TYPES (ICD-O 3)
Type here for selection ... [?]
SELECT CANCER LOCI
Type here for selection ... [?]
SELECT CANCER TYPES (NCIT NEOPLASMS)
Type here for selection ... [?]
ARRAY SERIES IDS
CITY (AND DISTANCE FROM IT)
km [?]

Query Database

4098 of 62104 cases matched the selection criteria.

¶ Subsets	¶ Samples	¶ Observations	¶ Frequency
9590/3: 9590/3: Malignant lymphoma, NOS	6	5	0.833
9052/3: 9052/3: Epithelioid mesothelioma, malignant	22	14	0.636
9050/3: 9050/3: Mesothelioma, malignant	87	55	0.632
9729/3: 9729/3: Precursor T-cell lymphoblastic lymphoma	30	14	0.467
8130/1: 8130/1: Papillary transitional cell neoplasm of low malignant potential	38	12	0.316
8250/3: 8250/3: Bronchiolo-alveolar adenocarcinoma, NOS	23	7	0.304
9801/3: 9801/3: Acute leukemia, NOS	10	3	0.300
9440/3: 9440/3: Glioblastoma, NOS	2047	607	0.297
8560/3: 8560/3: Adenosquamous carcinoma	21	6	0.286
8072/3: 8072/3: Squamous cell carcinoma, large cell, nonkeratinizing, NOS	96	27	0.281
9827/3: 9827/3: Adult T-cell leukemia/lymphoma (HTLV-1 positive)	179	49	0.274
9837/3: 9837/3: Precursor T-cell lymphoblastic leukemia	233	59	0.253
8500/2: 8500/2: Intraductal carcinoma, noninfiltrating, NOS	147	36	0.245
8430/3: 8430/3: Mucoepidermoid carcinoma	21	5	0.238
8811/3: 8811/3: Fibromyxosarcoma	43	10	0.233
8160/3: 8160/3: Cholangiocarcinoma	39	9	0.231
8012/3: 8012/3: Large cell carcinoma, NOS	48	11	0.229

ICD-O
Locus
NCIt
S
ER
C
?

Progenetix: Cancer Genome Profiles, Article Metrics, Epistemology, Resource Hub

cancer genome data @ progenetix.org

The Progenetix database provides an overview of copy number abnormalities in human cancer from currently **32317** array and chromosomal Comparative Genomic Hybridization (CGH) experiments, as well as Whole Genome or Whole Exome Sequencing (WGS, WES) studies. The data presented through Progenetix represents **364** different cancer types, according to the International classification of Diseases in Oncology (ICD-O).

Additionally, the website attempts to identify and present all publications (currently **2965** articles), referring to cancer genome profiling experiments. The database & software are developed by the group of Michael Baudis at the University of Zurich.

Publication data
2947 publications have been found.

New Search ...

Technique

- ACGH
- CCGH
- WES
- WGS

Publication	cCGH	aCGH	WES	WGS				
1-250	251-500	501-750	751-1000	1001-1250	1251-1500	1501-1750	2751-2947	all
Peng S, Dhur H, Armstrong B, Salgia B, et al. Integrated genomic analysis of survival outcomes in glioma. <i>Nature</i> . 2010;465(7294):916-920.	0	0	18	18				
Shi J, Hu X, Zhu B, Ravichandran S, Wang Y, et al. Somatic Genomics and Clinical Features of Lung Adenocarcinoma. <i>PLoS Med.</i> 13(12), 2016.	0	101	101	0				
Vanni I, Coco S, Bonfiglio S, Cittaro D, Gennari L, et al. Whole exome sequencing of independent breast carcinomas, and malignant peritoneal mesothelioma. <i>Cancer Res.</i> 2016;26(1):10-17.	0	0	3	0				
Lianos GD, Giannoutakis GK, Bali CD, Katsios C, Roukos DH, et al. Identification of novel genes by whole-exome sequencing can improve gastric cancer precision oncology. <i>Future Oncol.</i> 2016;22(3):261-270.	0	0	2	0				
Ferreira EN, Barros BD, de Souza JE, Almeida RV, Torrezan GT, Garcia S, Krepisch AC, et al. A genomic case study of desmoplastic small round cell tumor: comprehensive analysis reveals insights into potential therapeutic targets and development of a monitoring tool for a rare and... <i>Hum. Genomics</i> 10(1), 2016.	0	1	1	0				
Fiset PO, Fontebasso AM, De Jay N, Gayden T, Nikbakht H, Majewski J, Jabado N, et al. Longitudinal mutational analysis of a cerebellar pilocytic astrocytoma recurring as a ganglioglioma. <i>Pediatr Blood Cancer</i> , 2016.	0	0	4	0				
Bi WL, Horowitz P, Greenwald N, Abedalthagafi M, Agarwalla PK, Gibson WJ, Mei Y, et al. Landscape of genomic alterations in pituitary adenomas. <i>Clin. Cancer Res.</i> 2016;22(1):10-18.	0	0	42	0				
Zhao J, Xu W, He M, Zhang Z, Zeng S, Ma C, Sun Y, Xu C. Whole-exome sequencing of muscle-invasive bladder cancer identifies recurrent copy number variation in IPO11 and prognostic significance of importin-11 overexpression on poor... <i>Oncotarget</i> , 2016.	0	0	6	0				
Lips EH, Debipersad R, Scheereman CE, Mulder L, Sonke GS, van der Kolk LE, et al. BRCA1-mutated estrogen receptor positive breast cancer shows BRCAneSS, suggesting sensitivity to drugs targeting homologous recombination deficiency. <i>Clin. Cancer Res.</i> , 2016.	0	16	0	0				
Zhao F, Sucker A, Horn S, Heeke C, Bielefeld N, Schirr B, Bicker A, Lindemann M, et al. Melanoma Lesions Independently Acquire T-cell Resistance during Metastatic Latency. <i>Cancer Res.</i> 76(15), 2016.	0	5	0	0				

PROGENETIX NEWS

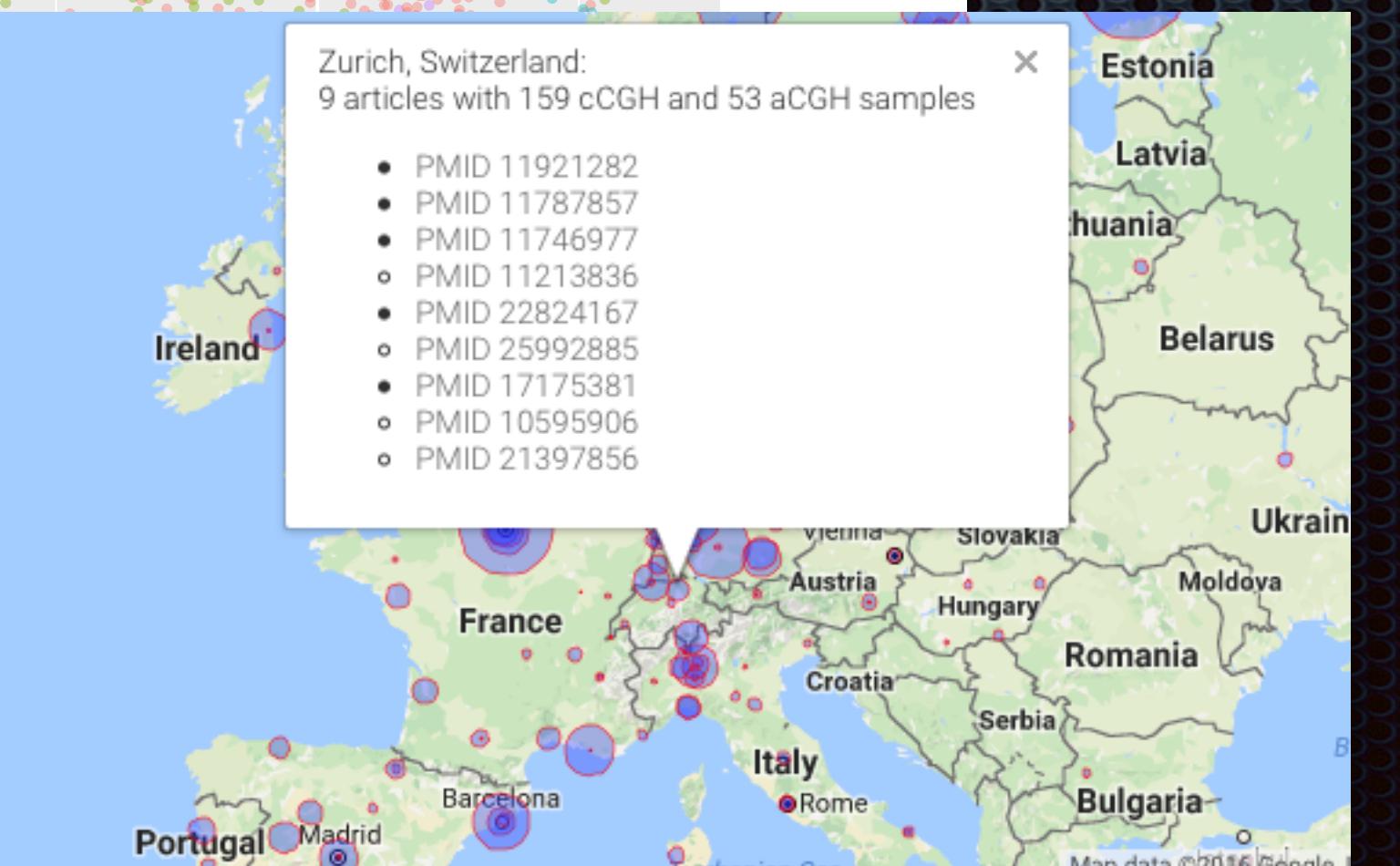
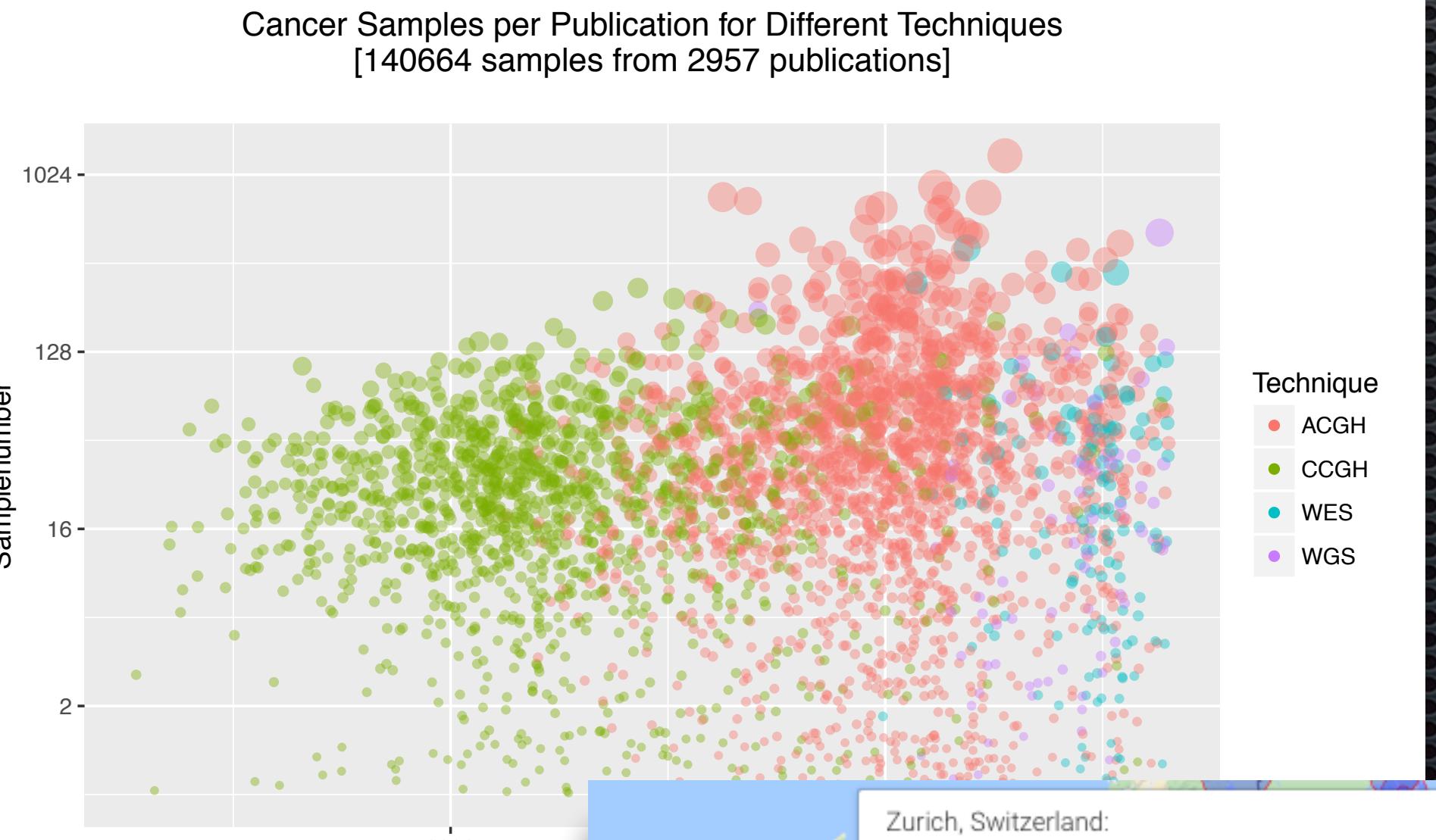
- 2016-08-03: SVG graphics
- 2016-05-17: Transitioning to Eu
- More news ...

RELATED PUBLICATIONS

Feel free to use the data and tools for academic research projects and other applications. If you have any questions, please contact Michael Baudis regarding a collaborative project or a special license.

© 2000 - 2016 Michael Baudis, refreshed Sat, 17 Dec 2016 16:53:31 GMT in 5.63 of the data presented nor the results achieved.

77.57.117.175



Progenetix: 12 years of oncogenomic data curation

Published online 12 November 2013 | *Nature Methods Research*, Vol. 12, Database issue | DOI:10.1038/d41573-013-0162

Authors
Hoeyang Cai^{1,*}, Nitin Kumar^{1,2}, Ni Ai^{1,2}, Saumya Gupta^{1,2}, Prisc Rath^{1,2} and Michael Baudis^{1,2}

Institutions
¹Institute of Molecular Life Sciences, University of Zurich, CH-8057 Zurich, Switzerland. ²Spiral Institute of Bioinformatics, University of Zurich, CH-8057 Zurich, Switzerland. ^{*}These authors contributed equally to this work.

Received August 21, 2013; **Revised** and **Accepted** October 21, 2013

ABSTRACT
DNA copy number aberrations (CNAs) can be found in the majority of cancer genomes and are crucial for understanding the molecular mechanisms of cancer development and progression. In its 12th release (v2.20), the Progenetix project (<http://www.progenetix.org>) has dedicated to provide the most comprehensive collection of CNAs from array comparative genomic hybridization (array CGH) and whole genome or whole exome sequencing (WGS/WES) studies. The data presented through Progenetix over the past 12 years were data curation efforts have resulted in the largest collection of cancer samples presented through Progenetix. In addition, the new user interface has been added in, particular, the gene options have been added in, particularly, the gene frequency and gene expression analysis, including various data representation options for primary and secondary analysis. This article reports recent improvements of the database in terms of user interface and other tools.

INTRODUCTION
DNA copy number aberrations (CNAs) are a form of genomic mutations found in the majority of individual cancer genomes. CNAs are often used to reveal both shared and distinct evolutionary processes in the same cancer type. Understanding the role and mechanism of CNAs in cancer requires analysis of oncogenes (2,7) and tumor-suppressor genes (8) in cancer cells through the different fluorescence intensity ratios after the detection of DNA labeled with different fluorescent dyes (13,14). For all types of hybridization targets, the same basic principle applies: a target DNA sequence is labeled and hybridized to the desired target DNA sequence. After hybridization, the fluorescence intensity ratio between reference DNA labeled with a different fluorescent dye and target DNA labeled with a specific probe genome (15,16). Single-color array experiments measure the relative fluorescence intensity ratios of each probe. Dual-color experiments measure the absolute fluorescence intensity ratios of each probe. Generally, the resolution of eCGH is limited to

RESEARCH ARTICLE
CDCOCA: A statistical method to define complexity dependence of co-occurring chromosomal aberrations

Nitin Kumar¹, Hubert Rehrauer¹, Hoeyang Cai¹, Michael Baudis¹

Abstract
Background: Copy number aberrations (CNAs) play a key role in cancer development and progression. Since most CNAs are shared among cancer types, testing methods for co-occurrence evaluation so far have not considered the overall complexity of the CNA patterns. Results: We propose a statistical method called CDCOCA to evaluate the complexity of CNAs and that these CNAs could be of pathogenic relevance for the respective cancer. We hypothesize that the complexity of CNAs is associated with the co-occurrence of CNAs. We validate our hypothesis with CDCOCA using chromosomal aberration data from 2957 publications. Conclusions: We have developed a method to detect associations of regional copy number anomalies in cancer. Our method is able to identify co-occurring CNAs in cancer, which may have negative impact on cancer prevention and treatment.

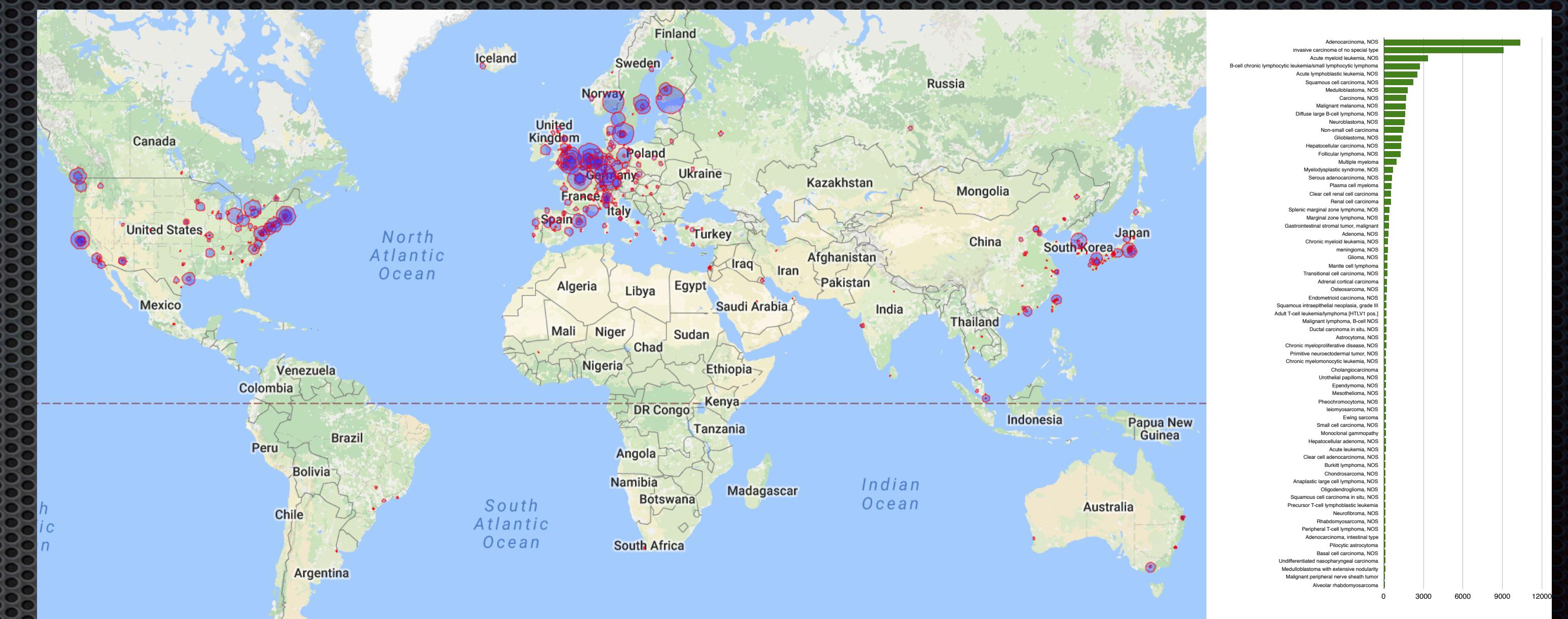
Background
Genetic alterations are an absolute requirement for malignant transformation. Both loss and gain of genetic alterations and order of events are important to sequential models, large-scale analyses of chromosomal aberrations (CNAs) have been identified in multiple cancer types. Recurrent CNAs have been identified in various cancer types. Chromosomal Comparative Genomic Hybridization (CGH) (11,12) is a generic wide CGH approach that can detect changes in the genome throughout the last two decades. Building on the reverse engineering of the genome-wide CGH approach, microarray CGH (13), genomic microarray technology (GCAL) (14), and next-generation sequencing (NGS) (15) have been developed to derive regional copy number anomalies in cancer. Compared with CGH, GCAL and NGS are more sensitive and require less sample material. Large data sets from copy number experiments are available for further analysis. However, the interpretation of these data is still challenging.

Competitive manuscript received: 06 June, 2013; **revision received:** 06 July, 2013; **accepted:** 06 July, 2013; **published online:** 12 November 2013. © 2013 Springer Nature Limited. All rights reserved. **Open Access** This article is published under the terms of Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0/>).

BioMed Central

Bias in Ascertainment / Background / Environment in Cancer Genome Studies

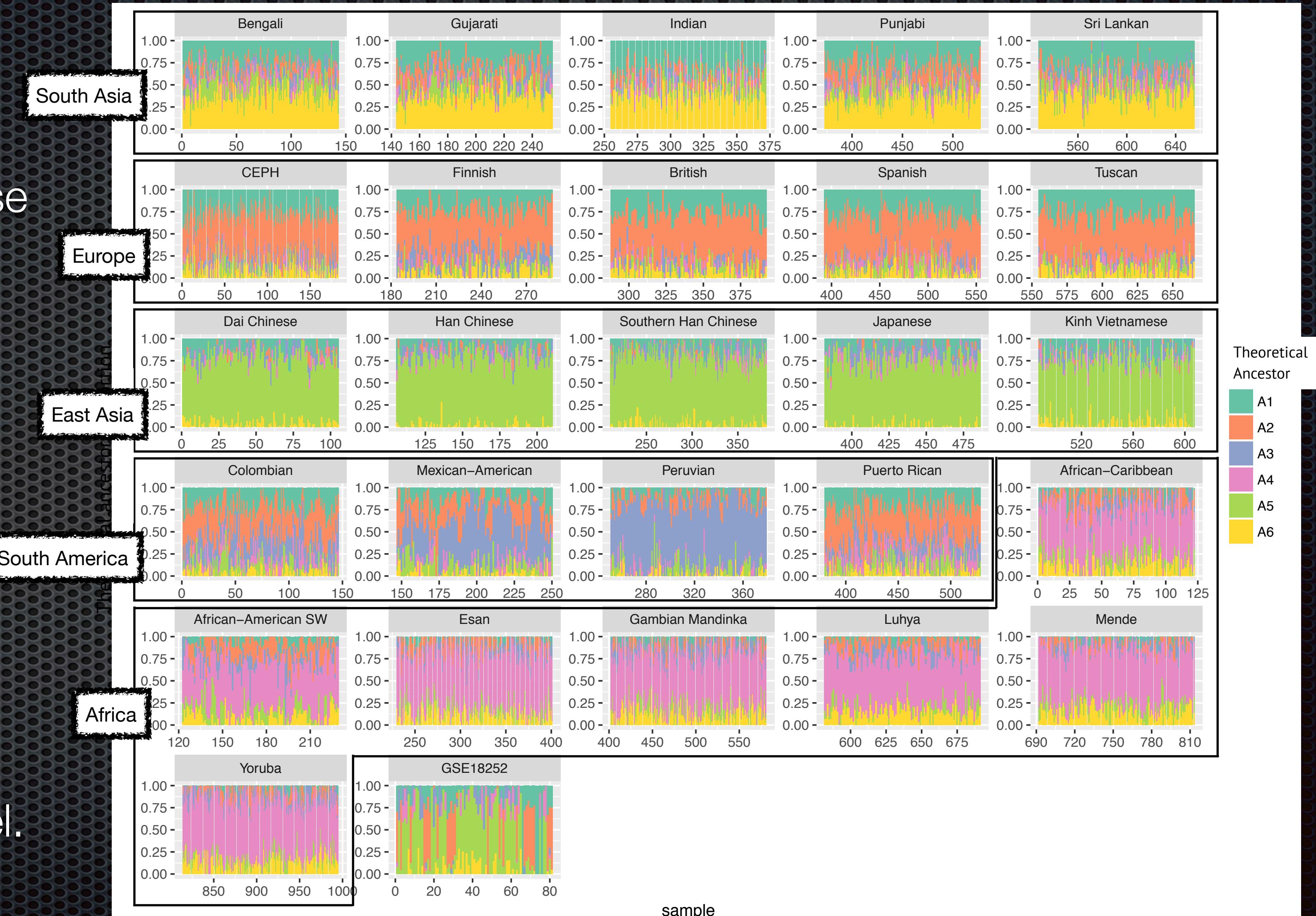
- the frequency of many genome variants depends on the genetic background
- cancer incidence & type can correlate to environmental factors
- geographic analysis can support interpretation and point to knowledge gaps



Geographic distribution of >140'000 cancer genome profiles reported in the literature. The numbers are derived from the 2947 publications registered in the Progenetix database.

Population stratification in cancer samples based on SNP array data

- 2504 genome profiles from 1000 Genome project phase 1 as reference
- 5 superpopulations: South Asia, Europe, South America, East Asia and Africa.
- SNP positions used in 9 Affymetrix SNP arrays are extracted to train a population admixture model.





University of
Zurich^{UZH}

Prof. Dr. Michael Baudis
Institute of Molecular Life Sciences
University of Zurich
SIB | Swiss Institute of Bioinformatics
Winterthurerstrasse 190
CH-8057 Zurich
Switzerland



Global Alliance
for Genomics & Health



arraymap.org

progenetix.org

info.baudisgroup.org

sib.swiss/baudis-michael

imls.uzh.ch/en/research/baudis